US006789070B1

(12) **United States Patent**
Willett et al.

(10) **Patent No.: US 6,789,070 B1**
(45) **Date of Patent: Sep. 7, 2004**

(54) **AUTOMATIC FEATURE SELECTION SYSTEM FOR DATA CONTAINING MISSING VALUES**

(75) Inventors: **Peter K. Willett**, Coventry, CT (US); **Robert S. Lynch, Jr.**, Groton, CT (US)

(73) Assignee: **The United States of America as represented by the Secretary of the Navy**, Washington, DC (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: 09/606,118

(22) Filed: Jun. 14, 2000

(51) Int. Cl.[7] **G06F 15/18**
(52) U.S. Cl. **706/20**; 706/19
(58) Field of Search 706/20, 19, 15; 702/181; 700/30, 89

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,109,438 A	*	4/1992	Alves et al.	382/243
5,765,127 A	*	6/1998	Nishiguchi et al.	704/208
5,999,893 A	*	12/1999	Lynch et al.	702/181
6,397,200 B1	*	5/2002	Lynch et al.	706/20

OTHER PUBLICATIONS

Lynch, Jr. et al., "Bayesian Classification and the Reduction of Irrelevant Features From Training Data", Proceedings of the 37 IEEE Conference on Decision and Control, pp. 1591–1592, vol. 2, Dec. 1998.*

Lynch, Jr. et al., "Testing the Statistical Similarity of Discrete Observations Using Dirichlet Priors", Proceedings of the IEEE International Symposium on Information Theory, p. 144, Aug. 1998.*

Lynch, Jr. et al., "A Bayesian Approach to the Missing Features Problem in Classification", Proceedings of the 38th Conference on Decision and Control, pp. 3663–3664, Dec. 1999.*

Voz et al., "Application of Suboptimal Bayesian Classification to Handwritten Numerals Recognition", IEE Workshop on Handwriting Analysis and Recognition: A European Perspective, pp. 9/1–9/8, Jul. 1994.*

Filip et al., "A Fixed–Rate Product Pyramid Vector Quantization Using a Bayesian Model", Gobal Telecommunications Conference 1992, vol. 1, pp. 240–244, Dec. 1992.*

Kontkanen et al., "Unsupervised Bayesian Visualization of High–Dimensional Data", ACM, 2000, Retrieved from the Internet: Http://www.cs.helsinki.fl/research/cosco.*

Zhang et al., "Mean–Gain–Shape Vector Quantization Using Counterpropagation Networks", Canadian Conference on Electrica and Computer Engineering, Sep. 1995, vol. 1, pp. 563–566.*

Baggenstoss, P., "Class–Specific Feature Sets in Classification", IEEE Transactions on Signal Processing, Dec. 1999, Vo. 47, No. 12.*

Basu et al., "Estimating the Number of Undetected Errors: Bayesian Model Selection", Proceedings of the 9th Intl Symposium o Software Reliability Engineering, Nov. 1998.*

(List continued on next page.)

Primary Examiner—Anthony Knight
Assistant Examiner—Kelvin Booker
(74) *Attorney, Agent, or Firm*—James M. Kasischke; Michael F. Oglo; Jean-Paul A. Nasser

(57) **ABSTRACT**

An automatic feature selection system for test data with data (including the test data and/or the training data containing missing values in order to improve classifier performance. The missing features for such data are selected in one of two ways: first approach assumes each missing feature is uniformly distributed over its range of values whereas in the second approach, the number of discrete levels for each feature is increased by one for the missing features. These two choices modify the Bayesian Data Reduction Algorithm accordingly used for the automatic feature selection.

8 Claims, 2 Drawing Sheets